Institutional Robotics

Porfírio Silva[1] and Pedro U. Lima[2]

[1] Philosophy Department, Faculdade de Letras, University of Lisbon
porfiriosilva@clix.pt
[2] Institute for Systems and Robotics, Instituto Superior Técnico,
Technical University of Lisbon
pal@isr.ist.utl.pt

Abstract. Pioneer approaches to Artificial Intelligence have traditionally neglected, in a chronological sequence, the agent body, the world where the agent is situated, and the other agents. With the advent of Collective Robotics approaches, important progresses were made toward embodying and situating the agents, together with the introduction of collective intelligence. However, the currently used models of social environments are still rather poor, jeopardizing the attempts of developing truly intelligent robot teams. In this paper, we propose a roadmap for a new approach to the design of multi-robot systems, mainly inspired by concepts from Institutional Economics, an alternative to mainstream neoclassical economic theory. Our approach intends to sophisticate the design of robot collectives by adding, to the currently popular emergentist view, the concepts of physically and socially bounded autonomy of cognitive agents, uncoupled interaction among them and deliberately set up coordination devices.

Key words: Collective Robotics, Institutional Economics, Institutional Robotics

1 Introduction

Three great neglects are at the heart of Good Old-Fashioned Artificial Intelligence: the neglect of the body, of the world, and of other agents. Collective Robotics is an important attempt to surpass these neglects: because it embodies intelligence in physical robots; because it places robots in physical environments largely natural; because it locates intelligence in the collective. Nevertheless, most multi-robot systems model extremely poor social environments. Our aim is to put forward a new conceptual approach to design control systems of artificial robotic societies. In Section 2 some weaknesses of popular guiding principles to collective systems design are identified. In Section 3 we look for inspiration coming from fields of sciences of the artificial other than robotics. In Section 4 we sketch out a new strategy to conceptualize multi-robot systems: *Institutional Robotics*, which takes institutions as the main tool of social life of robots with bounded rationality and bounded autonomy.

F. Almeida e Costa et al. (Eds.): ECAL 2007, LNAI 4648, pp. 595–604, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Emergence, Uncoupled Interaction, Bounded Autonomy, and Collective Inefficiency

Two out of the four design principles for collective systems suggested in [1:241–243] represent popular views among practitioners of AI and Robotics. According to the "level of abstraction principle", collective intelligence refers not only to groups of individuals, as in human societies, but equally "to any kind of assembly of similar agents", including groups of modules in modular robotic systems or organs that make up entire organisms. The "design for emergence principle" states that a desired functionality should not be directly programmed into a group of agents, but emerge from a set of simple rules of local interaction, relying on self-organizing processes. These two principles raise three questions.

First, exclusive focus on emergence and self-organization stems from the prominent role conferred to local interaction. No reason is given to ignore indirect or mediated interaction, which [2:14] considers characterized by properties such as name uncoupling (interacting entities do not have to know one another explicitly), space uncoupling and time uncoupling (they do not have neither to be at the same place nor to coexist at the same time). Communication is an example of such an indirect (not local) interaction.

Second, to put individuals in human societies on the same foot with organs or modules purports to ignore different degrees of autonomy enjoyed by a human lung and a human individual. According to [3:518], the following could be a working definition within robotics: "Autonomous agents operate under all reasonable conditions without recourse to an outside designer, operator or controller while handling unpredictable events in an environment or niche". However, for some philosophical perspectives this conception of autonomy would be unsatisfactory, because a truly autonomous agent must be capable of acting according to his own goals, while designers are the sole providers of goals to the robots. A concept of autonomy arising out of the ongoing attempt to maintain homeostasis could improve our understanding of autonomy and how goals become grounded in artificial creatures [3].

Whether full autonomy is attainable is a remaining question. A sharp negative answer to that question is offered by [4]. Autonomous agents must be capable of generating new goals as means for achieving existing goals of their own, but they are not necessarily self-sufficient. An agent depends on a resource where he needs it to perform some action to achieve one of his goals. There is also social dependence: an agent x depends on another agent y when, to achieve one of his goals, x needs an action of y. Two agents can be mutually dependent. Dependences imply interests: a world state that favours the achievement of an agent's goals is an interest of that agent. Dependence and interest relations are objective relations, holding whether an agent is aware of them or not. Limited autonomy of social agents comes also from influencing relations between them. By acquiring (true or false) beliefs about their interests agents can acquire goals. Now, an agent x can influence another agent y to adopt a goal according to x's needs, even if that goal is not an interest of y.

Within this approach, cognition does not preclude emergency. To form goals and establish plans to their achievement, agents must be cognitive. However, bounded rationality combines with bounded autonomy to give place to emergent phenomena: there are deliberately planned actions but they may produce unintended effects beyond reach of the agent's understanding or awareness.

Third, no reason is given to rule out coordination devices deliberately set up by agents in some multi-agent systems (legislation and organisations in human societies, for example). The remaining question is the desirability of that fact. Could we show that, at least in some situations, merely emergent processes may lead to inefficient solutions to collective problems? If so, we would have a hint on why multi-agent systems may need coordination devices. A set of experiences within MAS, reported in [5], advances our understanding of the problem. There, situations previously identified in experimental economics are simulated with a version of the Genetic Algorithm (GA). The GA population represents a collection of sets of rules associated with the set of actions available to agents; the fitness function for each agent maximizes his payments.

Co-ordination problem 1. A set of individuals, kept in isolation from one another, must choose one of 16 colours. Each participant choice will be rewarded in accordance with the rule: multiply a fixed amount of money by the number of players that have chosen the same colour. The experiment repeats a number of times. After each repetition, players are informed of frequencies and pay-offs by colour, so participants can change their choices next time, what they indeed do to maximize payments. Individual behaviours rapidly converge: the rule "choose colour x", where x is the most often selected, emerges as a shared convention. The "spontaneous order hypothesis" seems to work.

Co-ordination problem 2. A new experimental situation departs from the previous one in just one detail. The payoff to each individual now depends, not only on the frequency of the chosen colour, but also on an "intrinsic" characteristic of each colour, which remains unknown to players. For example, all other factors remaining equal, the choice of the colour number 16 pays 16 times more than colour number 1. The convergent choices of all participants to colour 16 is the most valuable situation to every participant, but that convergence is unlikely to occur in the absence of any opportunity to agree on a joint strategy. An initial accidental convergence to any colour creates an attractor capable of strengthen itself from repetition to repetition. Even if a participant has a god's eye view of the situation, any isolated option for the best theoretical option will neither improve the individual payoff nor move the collective dynamics towards a path conducive to a higher collective payoff. Self-organizing processes may lead to inefficient solutions for a collective problem. The "spontaneous order hypothesis" is in trouble, even with mere co-ordination problems, when the best for each individual is also the best for the collective (for other individuals). The situation gets worse with a "co-operation problem", when the best outcome for the collective and the best outcome for an individual don't coincide necessarily.

Co-operation problem. Now, the individuals must post a monetary contribution (from 0 to a predefined maximum) in an envelope and announce the amount contained in it. The sum of all the contributions is multiplied by a positive factor ('invested') and the resultant collective payoff is apportioned among the individuals. For each participant, its share of the collective payoff is proportional to the announced contribution, not to the posted contribution. As all participants know these rules, they realize that to maximize payoff an individual must contribute nothing and announce the maximum. So, it is with no surprise that, after some initial rounds, free-riding behaviour emerges: the posted contributions tend to zero while the announced contributions are kept close to the maximum. The group follows collectively a path that all of his members consider undesirable: soon there will be no more money to distribute.

This set of experiences suggests collective order does not always emerge from individual decisions alone. Coordination devices deliberately set up by agents could be useful.

3 Artefacts in Institutional Environments

The previous Section has shown that some concepts can add to emergentist views in order to sophisticate artificial collective systems design: physically and socially bounded autonomy of cognitive (not only reactive) agents; uncoupled interaction among them; deliberately set up coordination devices. How could we put all these concepts together? Social sciences' concepts have already inspired fields of sciences of the artificial other than robotics. Relying on some results of that cross-fertilization, we will arrive at the unifying concept of "institutional environment". It will later lead us to Institutional Robotics.

Epstein and Axtell argue that artificial society modelling can constitute a new kind of explanation of social phenomena [6:20]. Lansing [7] argues that the modelling of artificial societies can profit from a broad historical perspective of disputes among social scientists and philosophers on how to study social phenomena. To exemplify, he points out the parallel between some writing of Theodor Adorno on the positivist dispute in German sociology and the question that introduces [6]: "How does the heterogeneous micro-world of individual behaviors generate the global macroscopic regularities of the society?". This is a classical problem of the social sciences, the micro-macro link problem or the problem of social order. A number of researches take both perspectives together within Multi-agent systems (MAS) modelling. A few examples are: [8] studies norms as a possible solution to coordination problems; [9] suggests relaxing the assumption that coordination can be designed to perfection and importing conflict theories from sociology; [10] reviews trust and reputation models; within the framework of "Socionics" (a combination of sociology and computer science [11]), the Social Reputation approach [12] models reputation as an emergent mechanism of flexible self-regulation; [13] argues for using basic individual rights in MAS, combined with some argumentation mechanism.

Facing such a variety, how would we choose the most promising concept? Perhaps we need them all. "It does not seem possible to devise a coordination strategy that always works well under all circumstances; if such a strategy existed, our human societies could adopt it and replace the myriad coordination constructs we employ, like corporations, governments, markets, teams, committees, professional societies, mailing groups, etc." [14:14] So, we keep them all, and more – but we need an unifying concept to give the whole some consistence.

"Environment" is such a concept. [2] suggests the need to go deeper than the subjective view of MAS, where the environment is somehow just the sum of some data structures within agents. What we need to take into account is the active character of the environment: some of its processes can change its own state independently of the activity of any agent (a rolling ball that moves on); multiple agents acting in parallel can have effects any agent will find difficult to monitor (a river can be poisoned by a thousand people depositing a small portion of a toxic substance in the water, even if each individual portion is itself innocuous) [2:36]. Because there are lots of things in the world that are not inside the minds of the agents, an objective view of environment must deal with the system from an external point of view of the agents [15:128].

One can wonder if this can be relevant to robotics, where agents already behave sensing and acting in real (not just software) environments. We suggest the answer is affirmative. Dynamic environmental processes independent of agents' purposes and almost unpredictable aggregate effects of multiple simultaneous actions are not phenomena restricted to physical environments. Similar phenomena can occur in organizational environments: if nine out of ten of the clients of a bank decide to draw all their money at the same date, bankruptcy could be the unintended effect. And, most of the time, social environments in robotics are poorly modelled. So, the objective view of the environment could apply not only to physical features, but also to the social environment of the agents. We further suggest that both physical and social environments are populated with strange artefacts: artefacts with material and mental aspects. Let us see, following [16].

An artefact is something done by an agent to be used by another (or the same) agent. An artefact may not be an object: footprints left on a mined field for the followers are artefacts. Artefacts shaped for coordinating the agents' actions are coordination artefacts. Even single-agent actions can be coordinated actions if they contribute to solve an interference problem with other agents. Some artefacts have physical characteristics that represent opportunities and constraints which are sufficient conditions to enable a single-agent coordinated action, even if the agent doesn't recognize them (the wall of a house keeps people inside and outside separated). Sometimes, the agent must additionally recognize the opportunities and constraints of the artefact: sitting at a table with other people needs some knowledge ("not try to seat at a place already occupied").

More interesting artefacts are associated not only with physical but also with cognitive opportunities and constraints (deontic mediators, such as permissions and obligations). Recognizing all of those enables a single-agent coordinated action: a driver approaching a roundabout is obliged, only by physical properties of the artefact, to slow down and go right or left to proceed; traffic regulations add something more indicating which direction all drivers have to choose not to crash with others.

Furthermore, artefacts can be completely dematerialized. Such artefacts enable single-agent coordinated actions only by means of cognitive opportunities and constraints recognized by the acting agent. Social conventions and norms are relevant examples of the kind. A traffic convention to drive on the right works independently of any material device.

Consider now multi-agent coordinated actions. “There exist some artefacts such that the recognition of their use by an agent and the set of cognitive opportunities and constraints (deontic mediators) are necessary and sufficient conditions to enable a multiagent coordinated action” [16:320]. Institutions are of such a kind of artefacts. The definition takes institutional actions as multi-agent coordinated actions performed by a single-agent. How could this be? Because of a cognitive mediation intertwined with the agents’ behaviours. While traditional views on institutions take them as structured sets of rules and conventions, in [16] the basic coordination artefact is the institutional role played by an agent with the permission of others. A group of agents recognizes that an agent (Paul) plays a role (priest) and so believes he has the artificial power of doing a multi-agent coordinated action (the marriage of John and Mary). Both recognition and belief are intertwined with the behaviour of treating Paul as a priest and treating John and Mary, from some point in time on, as a married couple.

The single-agent action of an agent playing a role is the vehicle action for a collective action, like flipping the switch is the vehicle action for the supra-action of turning the light on. In this context, the agent relies on some external aspects of the world (the functioning of the electrical circuit). To get John and Mary married the priest must perform a certain set of bodily movements counting as marrying. That set of movements is the vehicle action for the supra-action of marrying John and Mary. Again, the collective of agents rely on some external aspects of the world: the institutional context [16:312,320–321].

So, we have got our unifying concept: institutional environments populated with a special kind of artefacts.

4 Institutional Robotics

With the “institutional environment” concept as a starting point, in this Section we sketch out a new strategy to conceptualize multi-robot systems. Some global inspiration comes from Institutional Economics [17], an alternative to mainstream neoclassical economic theory. “Market-based multi-robot coordination” is a previous example of importing some Economics’ views into Robotics [18]. We do the same, but with different assumptions.

(1) The control system for a robotic collective is a network of institutions. All institutions exist as means for some activity of some set of robots. As a principle, institutions are generic: they are not designed to any specific set of robots.

(2) Institutions are coordination artefacts and come in many forms: organizations, teams, hierarchies, conventions, norms, roles played by some robots, behavioural routines, stereotyped ways of sensing and interpret certain situations, material artefacts, some material organization of the world. A particular institution can be a composite of several institutional forms.

(3) Institutions can be mental constructs. An example of a “mental institution” is a program to control a sequence of operations.

(4) Institutions can be material objects functioning exclusively by means of its physical characteristics given the physical characteristics of the robots (a wall separating two buildings effectively implements the prohibition of visiting neighbours if the robots are not able to climb it). Some rules (or other kinds of mental constructs) can be associated to a material object to create a more sophisticated institution (a wall separating two countries is taken as a border; there are some doors in the wall to let robots cross the border; some regulations apply to crossing the border).

(5) The boundaries between institutional and purely physical aspects of the world are not sharp. Not all material objects are institutions. If the wall separating buildings is seen as just an element of the physical world, some robots gaining access to opposite building with newly acquired tools or physical capabilities will not be minded as a breach of a prohibition. However, modifications of the material world creating new possibilities of interaction can become institutional issues. If the collective prefers to preserve the previous situation of separated buildings, the new capability of the robots to climb the wall could give place to new regulations. Material objects are devices for institutions when they implement some aspect of the collective order. The continuing existence of a material object can be uncoupled from the continuing existence of the institutional device it implements (the wall could be demolished without eliminating the border; the border can be eliminated without demolishing the wall). So, a material leftover of a discarded institution can last as an obstacle in the world.

(6) Enforcement mechanisms can be associated with institutions to prevent (or to redress negative effects of) violation. Examples are fines and reputation.

(7) The institutional environment at any point in the history of a collective is always a mix of inherited and newly adopted forms. So, the designer of a robotic collective must shape the first version of any system of institutional robotics. However, that first institutional setup must be neither purely centralized, nor fully decentralized, nor purely distributed. That means the following. Not all robots are equal in power: neither all agents have the same computational power, nor all access the same information, nor all are allowed to take decision on all domains. There are some hierarchical relations among robots: for any robot, access to information and permission to act are bounded by decisions of some others. However, different hierarchies apply to different issues and the same robot can be on top of one hierarchy and at bottom of others. Some robots, by virtue of one-to-one relationships not framed by any hierarchy, are able to establish short cuts to and influence top level decision makers that would otherwise be beyond reach. There is neither a single robot nor a small group of robots in charge of all collective decisions all the time. Although, some kind of elitism is possible: different (eventually partially overlapping) groups of robots share the ruling over different domains of decision. Elite must eventually be renewed: robots can be removed from power, robots can access power.

(8) Agents are robots, hardware/software “creatures”, operating on real physical environments. Robots are able to modify at some extent the material organization of their physical world.

(9)The continuing functioning of any robot depends on some material condition (available energy, for example). Whatever set of tasks a robot has to fulfil, some of them must be related to survival. There could be some institutions in charge of managing life conditions for all or some robots.

(10) All robots have built-in reactive behaviours, routines, and deliberative competences. Robots have partial models of themselves (they know some, but not all, of their internal mechanisms). Some of the internal mechanisms known by the robots can be accessed and modified by themselves.

(11) Every agent is created with links to some subset of the institutional network in existence within the collective. (Nobody is born alone in the wild). At some extent agents are free to join and to disconnect themselves from institutions. However, under certain circumstances, some institutions could be made compulsory for every agent or for some subset of all agents. Some institutions can filter access, either according to some objective rules or according to the will of those already connected. Disconnecting from an institution prevents the access to resources under control of it, as well as the participation in decision making processes taking place within it.

(12) Each robot has a specific individual identification (a name). All robots are able to identify, if not all, at least some others by their names.

(13) Any agent disconnected from all institutions will be treated by other agents as just an aspect of the material world. To recover from that extreme situation and get connected again to the institutional network an agent must be helped by some benevolent agent.

(14) World models are a special kind of institution. Being created with pre-established links to some institutions, any robot is endowed with some partial world models. World models can be specific to aspects of the physical world, specific to aspects of the social world or combine some aspects of both. None of the robots is endowed with a complete model of the world (except if gods are allowed). Inconsistencies between partial world models of one robot are allowed.

(15) There will be some process of collective world modelling. For example, a shared model of physical world can result from co-operative perception (sensor fusion [19:17–22]: merging sensor data from sensors spread over different robots and applying confidence criteria to weight their contribution to an unified picture of some aspect of the environment).

(16) The functioning of the sensorial apparatus of the agents can be modulated by their links to some institutions (adhering to an institution can augment the power or distort the functioning of some sensor). Institutional links can also modify the access to pieces of information available at collective level.

(17) From the point of view of an external observer the world model of a robot can be inaccurate. Inaccuracies can result from objective factors, intrinsic to the robotic platform (like sensors' limitations) or extrinsic (inappropriate vantage points to obtain data from some regions of the environment). Other inaccuracies can result from subjective factors: robots can have "opinions" and "ideologies".

(18) An "opinions" is an individual deviation from world models provided by institutions. (Even if a specific "opinion" of an individual agent is objective

knowledge gathered by virtue of some privileged vantage point, in such a manner that an external observer would prefer to rely on that opinion instead of accepting the "common sense", that means nothing to other agents, as long as they are deprived of that gods' view). By virtue of bearing an "opinion" the behaviour of a robot can be modified.

(19) An "ideology" is a set of "opinions" shared by a subset of all agents. Its acceptance among agents largely overlaps with sets of agents linked to some subset of the institutional network. An "ideology" can be "offered" by an institution to any agent prone to adhere or be a condition for adhesion. An "ideology" can result from a modification of the sensor fusion process (modification of the criteria to weight different individual contributions, for example). "Ideologies" can be about the physical or the social world. Modifying the perception of the agents and their behaviours, "ideologies" can affect the functioning of institutions in many ways: for example providing alternative stereotyped ways of sensing certain situations ("ignore such and such data streams") or undermining mechanisms of social control ("break that rule and we will pay the fine for you with a prize").

(20) Decision-making processes are a special kind of institution. Many aspects of collective dynamic can be subject to co-operative decision-making [19:34–46].

(21) Institutional building is a special issue for decision-making processes: "constitutional rules" for the functioning of some already existing institutions can be deliberated by the robots themselves; robots can deliberately set up new institutions or abandon old ones. Some institutions will have specific mechanisms to facilitate institutional building.

(22) Institutional building is fuelled by "institutional imagination": robots can conceive alternative institutions, or alternative constitutional rules to existing institutions, not to implement them at short term, but as "thought experiments". Results of those thought experiments can be put forward to specific institutional building mechanisms.

(23) The functioning of an institution can be modified, not by deliberative means, but by accumulating small modifications initiated by some robots and not opposed by others.

(24) An institution fade away when none agent is anymore linked to it. Robots can have memories of old institutions and reintroduce them in the future.

5 Conclusion

This paper suggested a new strategy to conceptualize multi-robot systems: the Institutional Robotics, which takes institutions as the main tool of social life of robots with bounded rationality and bounded autonomy. We have plans to set up a working group consisting of a team of people with a multidisciplinary background (e.g., philosophy, cognitive sciences, biology, computer engineering, artificial intelligence, systems and control engineering) to work on it, including further brainstorming, concepts refinement and actual implementation.

References

1. Pfeifer, R., Bongard, J.: How the Body Shapes the Way We Think. The MIT Press, Cambridge (2007)
2. Weyns, D., Parunak, H.: v. In: Weyns, D., Parunak, H.V.D., Michel, F. (eds.) E4MAS 2004. LNCS (LNAI), vol. 3374, pp. 1–47. Springer, Heidelberg (2005)
3. Haselager, W.F.G.: Robotics, philosophy and the problems of autonomy. Pragmatics & Cognition 13(3), 515–532 (2005)
4. Conte, R., Castelfranchi, C.: Cognitive and Social Action. The University College London Press, London (1995)
5. Castro Caldas, J., Coelho, H.: The Origin of Institutions: socio-economic processes, choice, norms and conventions. Journal of Artificial Societies and Social Simulation 2(2) (1999), `http://jasss.soc.surrey.ac.uk/2/2/1.html`
6. Epstein, J.M., Axtell, R.: Growing Artificial Societies: Social Science from the Bottom Up. Brookings Institution Press, Washington (1996)
7. Lansing, J.S.: Artificial Societies" and the Social Sciences. Artificial Life 8, 279–292 (2002)
8. Hexmoor, H., Venkata, S.G., Hayes, R.: Modelling social norms in multiagent systems. Journal of Experimental and Theoretical Artificial Intelligence 18(1), 49–71 (2006)
9. Malsch, T., Weiß, G.: Conflicts in social theory and multiagent systems: on importing sociological insights into distributed AI. In: Tessier, C., Chaudron, L., Müller, H.-J. (eds.) Conflicting Agents. Conflict Management in Multi-Agent Systems, pp. 111–149. Kluwer Academic Publishers, Dordrecht (2000)
10. Sabater, J., Sierra, C.: Review on Computational Trust and Reputation Models. Artificial Intelligence Review 24(1), 33–60 (2005)
11. Malsch, T., Schulz-Schaeffer, I.: Socionics: Sociological Concepts for Social Systems of Artificial (and Human) Agents. Journal of Artificial Societies and Social Simulation 10(1) (2007), `http://jasss.soc.surrey.ac.uk/10/1/11.html`
12. Hahn, C., Fley, B., Florian, M., Spresny, D., Fischer, K.: Social Reputation: a Mechanism for Flexible Self-Regulation of Multiagent Systems. Journal of Artificial Societies and Social Simulation 10(1) (2007)
13. Alonso, E.: Rights and Argumentation in Open Multi-Agent Systems. Artificial Intelligence Review 21(1), 3–24 (2004)
14. Durfee, E.H.: Challenges to Scaling Up Agent Coordination Strategies. In: Wagner, T.A. (ed.) An Application Science for Multi-Agent Systems, pp. 113–132. Kluwer Academic Publishers, Dordrecht (2004)
15. Weyns, D., Schumacher, M., Ricci, A., Viroli, M., Holvoet, T.: Environments in Multiagent Systems. The Knowledge Engineer Review 20(2), 127–141 (2005)
16. Tummolini, L., Castelfranchi, C.: The cognitive and behavioral mediation of institutions: Towards an account of institutional actions. Cognitive Systems Research 7(2-3), 307–323 (2006)
17. Hodgson, G.M.: Economics and Institutions: A Manifesto for a Modern Institutional Economics. Polity Press, Cambridge (1988)
18. Dias, M.B., Zlot, R.M., Kalra, N., Stentz, A.: Market-based multirobot coordination: a survey and analysis. Proceedings of the IEEE 94(7), 1257–1270 (2006)
19. Lima, P.U., Custódio, L.M.: Multi-Robot Systems. In: Innovations in Robot Mobility and Control. Studies in Computational Intelligence, vol. 8, pp. 1–64. Springer, Heidelberg (2005)